Eesti suulise kõne korpuse automaatne pindsüntaktiline analüüs

Kaili Müürisep, Helen Nigol, Heli Uibo
Tartu Ülikool

1. Sissejuhatus

Suulise kõne korpuse käsitsi süntaktiline märgendamine on keerukas ja aeganõudev protsess. Et seda lihtsustada, võtsime kasutusele eesti keele kitsenduste grammatika (ESTKG) analüsaatori (Roosmaa jt 2003), mis oli algselt loodud kirjaliku keele tekstide analüüsimiseks.

Eesti keele suulise kõne korpuse loomine algas 1997. a (Hennoste jt 2000). Hetkel on korpus 700000-sõnaline. Korpuses on nii argi- kui ka avaliku suhtluse tekste, nii spontaanset kui ka ettevalmistatud kõnet, nii dialooge kui ka monolooge. Korpuse litereerimisel kasutatakse konversatsioonianalüüsi transkriptsiooni. Korpuse osa tekste on morfoloogiliselt analüüsitud, osa on ka morfoloogiliselt ühestatud ja osa on märgendatud dialoogiaktide märgendusega (Hennoste jt 2003).

Oma eksperimentides kasutasime morfoloogiliselt ühestatud tekste. Süntaksi-analüsaatori kohandamiseks suulise kõne analüüsimiseks tuli muuta osalausepiiride reegleid, parandada mitmeid süntaktilisi kitsendusi ning võtta kasutusele paar uut märgendit.

2. Eesti keele kitsenduste grammatika süntaksianalüsaator

Eesti keele kitsenduste grammatika analüsaator töötati välja 1996-2001 Tartu ülikoolis. Süntaktilise analüüsi protsess on selles jaotatud kaheks osaks. Morfoloogiline ühestaja tegeleb kontekstiinfo põhjal morfoloogiliselt mitmese analüüsiga sõnavormile õige morfoloogilise kirjelduse väljavalimisega, süntaksianalüsaator leiab sõnavormi süntaktilise funktsiooni lauses. Meie analüsaator põhineb kitsenduste grammatikal (Karlsson jt 1995), mis on loomult reduktsionistlik, s.o analüüsi alguses lisatakse igale sõnavormile kõik võimalikud analüüsivariandid ja seejärel hakatakse konteksti mittesobivaid eemaldama. Eemaldamine toimub vastavalt kitsenduste grammatika reeglitele ehk kitsendustele, mis igaüks esitab mõnda spetsiifilist keelereeglilaadset fakti. Üldisem grammatikareegel kujuneb alles nende koosmõjust. ESTKG-s on hetkel

1118 süntaktiliste märgendite eemaldamise reeglit.

Ideaaljuhul jääb analüüsi lõppedes igale sõnavormile üks süntaktiline märgend. Kui sõnal võib olla lauses mitu funktsiooni, antakse need kõik. Mitme märgendiga jäävad ka sõnad, mida analüsaator pole suutnud lõpuni ühestada. Grammatikareeglid on kirjutatud nii, et pigem jäetakse sõna mitme analüüsiga, kui eemaldatakse korrektne märgend.

ESTKG-s märgendatavad süntaktilised funktsioonid vastavad enam-vähem standardses eesti keele grammatikas (Erelt jt 1993) eristatavatele süntaktilistele funktsioonidele.

Öeldise märgendid eristavad finiitset ja infiniitset öeldist ning eraldi märgendid on põhiverbile (@+FMV, @-FMV) ja abi- ning modaalverbidele (@+FCV, @-FCV). Fraasi põhjadest märgendatakse alust, sihitist, öeldistäidet, määrust (vastavalt @SUBJ, @OBJ, @PRD, @ADVL). Laiendite märgendid näitavad põhja leidumise suunda, kuid ei viidata ühelegi sõnale konkreetselt. See tähendab, et on eraldi märgendid ees- ja järeltäienditele (@NN>, @<NN jt), eessõna ja tagasõna laienditele (@<P, @P>) ning kvantori ees- ja järellaienditele (@Q>, @<Q). Täienditest eristatakse omadus-, määr-, kaas- ja nimisõnalisi täiendeid ning partitsiipe ja infinitiivseid verbivorme täiendina.

Näide automaatselt analüüsitud tekstist on toodud joonisel 1. Iga sõnavormi all on antud selle tüvi ja lõpp, morfoloogiline kirjeldus kaldkriipsude vahel ning süntaktiline märgend (algab @-sümboliga).

Kahjuks ei ole võimalik kõiki sõnu automaatselt ühestada, ligikaudu iga kümnes sõna jääb mitme märgendiga. Ilukirjandusliku teksti analüüsi tulemused on toodud tabelis 1, kus teises veerus on toodud tulemused, kui tekst oli eelnevalt käsitsi morfoloogiliselt ühestatud, ning kolmandas veerus täisautomaatse analüüsi tulemused. Saagis näitab, mitu protsenti sõnadest on õige märgendiga, pööramata tähelepanu sellele, kas sõna on ühene või mitte. Täpsus näitab, mitu protsenti kõigist märgenditest on omal õigel kohal ehk siis leitud korrektsete märgendite arvu suhet kõigi leitud märgendite arvu. Ühesus näitab, mitu protsenti sõnadest on ühese analüüsiga.

	Käsitsi ühestatud	Automaatselt ühestatud
Saagis	98,53%	96,41%
Täpsus	87,57%	78,09%
Ühesus	89,54%	82,70%

Tabel 1. Analüüsi tulemused.

```
Oli
   ole+i //_V_ main indic impf ps3 sg ps af #cap #Intr // **CLB @+FMV
päikesepaisteline
   päikese_paiste=line+0 //_A_ pos sg nom #line // @AN>
hommikupoolik
   hommiku_poolik+0 //_S_ com sg nom // @SUBJ
$,
   , //_Z_ Com //
mustad
   must+d //_A_ pos pl nom // **CLB @AN>
laigud
   laik+d //_S_ com pl nom // @SUBJ
aurasid
   aura+sid //_V_ main indic impf ps3 pl ps af #FinV // @+FMV
keset
   keset+0 //_K_ pre #part // @ADVL
määrdunud
   määrdu=nud+0 //_A_ pos #nud partic // @VN>
lund
   lumi+0 //_S_ com sg part // @<P
$.
   . //_Z_ Fst //
```

Joonis 1. Näide automaatselt süntaktiliselt analüüsitud kirjaliku keele tekstist.

3. Süntaksianalüsaatori kohandamine suulisele keelele

Et kohandada kirjaliku keele analüsaatorit suulise keele analüüsiks, tuli lisada uusi märgendeid ja reegleid osalausepiiride tuvastamiseks. Samuti tuli muuta mitmeid süntaktilisi kitsendusi.

3.1. Uued märgendid

Uute märgenditena võeti kasutusele @B partikli tähistamiseks ja @T tundmatu sõna märkimiseks.

Suulises kõnes esineb palju partikleid ja neid vaadatakse kui eraldi sõnaliiki. Partiklid on muutumatud omaette tüvega sõnad, mis võivad esineda ka kirjalikus keeles (*siis*, *jah*) või olla kirjakeele sõnade häälduslikud variandid (*sis*, *kule*), osa partikleid on aga häälikuühendid, mis paiknevad foneetiliselt ja fonotaktiliselt häälitsuse piirimail (*öök*, *phtüi*, *mhmh*). Süntaktiliselt võivad need üksused moodustada vestluses terve kõnevooru, seega ka süntaktilise üksuse (*mhmh*). Kui partiklid kuuluvad mingisse intonatsiooniliselt terviklikku pikemasse üksusesse, nt lausesse, siis ei kuulu nad lause grammatilisse struktuuri. Nad ei seostu mingi kindla sõnaklassiga lauses, kuid võivad töötada kogu lause juurde kuuluvate lauselaienditena (Hennoste 2002).

Kõnes esineb sageli grammatiliselt ebakorrektseid või poolikuid lauseid, samuti leidub sõnu, mille lausumine on poole pealt katkestatud ning seetõttu on nad automaatsel morfoloogilisel analüüsil märgendatud kui tundmatud sõnad. Sellistes lausetes on mõnede sõnade süntaktilise funktsiooni määramine ka lingvisti poolt võimatu ning sellised sõnad on treening- ja testkorpuses märgendatud kui tundmatu süntaktilise funktsiooniga sõnad - @T. Näites 1 ei ole võimalik kahte viimast sõna süntaktiliselt analüüsida, sest lause on lõpetamata.

(1) A: kui (@J) sa (@SUBJ) võtame (@+FMV) mingisuguse (@NN>) asja (@OBJ) kuigi (@J) seda (@T) see (@T)

3.2. Süntaktilise analüüsi aken

Kirjalikus tekstis on üheks analüüsiühikuks ehk süntaktiliseks aknaks, mille ulatuses vaadatakse sõna süntaktilise funktsiooni määramisel konteksti, lause. Kirjaliku keele lause on autori tahtel märgistatud punktuatsioonimärkidega. Suulisele kõnele on iseloomulik, et seda on väga raske lauseteks jagada. Suulises kõnes on lauselõpu-tunnusteks pausid ja intonatsiooni langused või tõusud. Litereerimisel on need ka erisümbolitega märgistatud. Kõnevool jagatakse intonatsioonilisteks üksusteks, mitte grammatilisteks üksusteks. Selliseid põhiüksusi on kaks: a) üksus, mida võib nimetada lausungiks, selle lõpus on selgelt langev intonatsioon, mis osutab lõpetatusele ning mida märgitakse punktiga; b) lausungid jagunevad intonatsiooniliselt osadeks, mille lõpus intonatsioon langeb, kuid vähem kui lausungi lõpus (poollangev intonatsioon). Selline intonatsioon osutab, et tegu on piiriga, kuid üksus ei lõpe. Seda märgitakse komaga. Lisaks kasutatakse tõusva intonatsiooniga lõppeva üksuse lõpus küsimärki.

Transkriptsioonimärgenduse täpsemal uurimisel ilmnes siiski, et lausungiteks jagamine polnud piisavalt täpne. Sageli tekkis intonatsioonilangus ka keset süntaktilist lauset. Näiteks: *ma tõstan kartulid ära. ja sousti ka.* Seepärast otsustasime käsitleda süntaktilise aknana ühte kõnevooru ning vaadelda punktidega eraldatud üksusi kui koordineeritud lauseosi.

Ilmselt sõltub lause piiri määramine siiski teksti liigist, monoloogide puhul on punktide kasutamine lauselõputunnusena igati õigustatud.

Samas võib tekstist leida ka näiteid, kus kõnevoor jagab lause pooleks (vahele rääkimine). Näites 2 tuleb sõnavorm *auto* esimesel korral märgendada kui tundmatu süntaktilise funktsiooniga sõna, sest vooru sees ei ole võimalik selle funktsiooni

määrata.

(2) A: see oli kõik ee ausatel eesmärkidel et perele auto
B: ei no loomulikult
A: saada. aga lissalt mai saand seda autot

3.3. Osalausepiiride määramise reeglid

Kirjaliku keele osalausepiirid määratakse sidesõnade, kirjavahemärkide ja verbide põhjal. Osalause esimesele sõnale lisatakse märgend CLB. Osalausepiiride määramise põhireegel on järgmine: kui sõnale eelneb kirjavahemärk ja/või sõna ise on sidesõna ning vasakul ja paremal pool seda sõna leidub verbi pöördeline vorm, siis see sõna on osalause esimene sõna. See reegel võib mõnede tingimuste osas varieeruda.

Koma või rinnastavate sidesõnade *ja*, *ning*, *või*, *ega*, *ehk* abil võib eraldada mitte ainult osalauseid, vaid ka koondlause korduvaid liikmeid. Seda, millise eraldajaga just konkreetsel juhul on tegu, on ilma süntaktilist informatsiooni teadmata raske otsustada, eriti veel juhul, kui antud sõna lähemas kontekstis ei leidu verbe. Seepärast lisatakse nendele sõnadele üksnes oletatava osalause tunnus CLB-C.

Kirjaliku eesti keele kitsenduste grammatikas on 47 osalausepiiride määramise reeglit, paljud neist on väga spetsiifiliste juhtude jaoks.

Kõik kirjaliku keele osalausepiiride määramise reeglid tuli ümber vaadata, sest kirjavahemärkide tähendus on suulise kõne tekstides erinev. Uutes reeglites kasutati intonatsioonimärke ja partikleid (*noh*, pausi täitjad *aa* ja *ee* jt). Punkti loetakse osalausepiiri kindlaks lõputunnuseks. Partiklit loetakse eraldajaks, kui kummalgi pool kontekstis leidub finiitseid verbivorme.

Reeglid püüavad leida valestarte ja märgendada neid osalausepiiri märgenditega, kuid see õnnestub ainult juhul, kui valestart sisaldab verbi (näide 3).

(3) mul (CLB) on kassetil (CLB-C) oleks ruumipuudus tekkinud

Väga raske on tuvastada sisemiste osalausete lõppu ning need on peamised vigade tekkimise kohad. Näites 4 on tuvastamata osalauselõpp tähistatud tärniga.

(4) kuna (CLB) ta tundus mulle esmakuulamisel või noh algul kui (CLB) ma kuulasin (*) kuidagi liiga afišlik või noh noh äraleierdatud.

Kirjaliku keele tekstis oleks see osalausepiir tähistatud komaga.

Suulise kõne analüüsi grammatika sisaldab 21 osalausepiiride määramise reeglit, mida on oluliselt vähem kui kirjalikus keeles. Seda võib seletada sellega, et kirjaliku keele

grammatika oli kohandatud analüüsima ka juriidilisi tekste, mille osalausepiiride määramisel peab arvestama paljude spetsiifiliste juhtudega (loetelud, paragrahvid, sulgudes tekst).

3.4. Parandused süntaktilistes kitsendustes

Nagu eelpool mainitud, koosneb algne kirjaliku keele grammatika 1118 reeglist. Neid reegleid rakendati 2200-sõnalisele eelnevalt süntaktiliselt märgendatud treening-korpusele (argivestlus) ning saadud vigade analüüsimisel parandati või muudeti reegleid. Enamasti tuli muuta reeglite kontekstipiiranguid. ESTKG reegleid võib rakendada kolmes režiimis: a) kontekstitingimused kehtivad kogu lause ulatuses, b) kontekstitingimused kehtivad ainult kindlate osalausepiiride vahel (CLB), c) kontekstitingimused kehtivad kõigi osalausepiiride sees (CLB ja CLB-C). Enamasti oli vigade põhjuseks asjaolu, et reeglid arvestasid liiga kauget konteksti, mis asus väljaspool tegelikku osalauset. Selle parandamiseks tuli muuta vaid konteksti kontrollimise režiimi tunnust. Reegel joonisel 2 eemaldab aluse märgendi, kui osalauses on ainsuse 1. pöördes verb ja sõna ise on osastavas. Esimene on algne reegel, teine modifitseeritud.

(5)
```
(@w =s0 (@SUBJ) (0 Par)(*-1C Sg1) **CLB)
  (@w =s0 (@SUBJ) (0 Par)(*-1C Sg1) **CLB-C)
```

Ka tuli üle vaadata kõik kirjavahemärke kasutavad reeglid ja võimalusel täiendada neid tingimustega, kui kirjavahemärke lauses ei leidu.

Samuti on sõnavara kasutus suulises keeles erinev. Näiteks tuli arvestada, et relatiivpronoomenit *mis* võidakse kasutada küsilauses küsisõna *kas* asemel või võrdlustes *nagu* ja *kui* asemel.

Reeglite muutmine lõpetati, kui vigade protsent vähenes 7,5-lt 3-le.

4. Hindamine

4.1. Testkorpuse kirjeldus

Analüsaatori töö hindamiseks loodi käsitsi süntaktiliselt märgendatud testkorpus[1], mis koosnes 2543 sõnast. Testkorpus koosnes argivestlustest, milles oli nii pikemaid

[1] http://www.ut.ee/~kaili/Korpus/Spoken

jutustavaid dialooge kui ka lühikeste remarkidena dialooge. Korpus loodi sel viisil, et parandati analüsaatori poolt põhjustatud vead käsitsi ära. Parandajaks oli üks inimene.
Sellisel hindamisel on nõrgad kohad, mis on ka autoritele teada:
1. Korpus on liiga väike ega hõlma suulise kõne kõiki tahke.
2. Automaatselt analüüsitud korpust parandades võivad jääda mitmed vead märkamata, sest esialgsel vaatamisel tundub märgend olema korrektne ning inimene ei süüvi peensustesse.
3. Korpust peaks vähemalt esialgsel etapil märgendama mitu inimest, mis tagaks kõigi vigade avastamise ja järjekindlama märgenduse jälgimise.

4.2. Tulemused

Süntaksianalüsaatori väljundit võrreldi käsitsi märgendatud korpusega ning saadi järgmised tulemused (kirjaliku keele andmed on toodud sulgudes):
- sõnade arv korpuses: 2543
- saagis: 97.3% (98.5%)
- täpsus: 89.2% (87.5%)
- ühesus: 91.5% (89.5%)

4.3. Vigade tüübid

Vead võib jagada järgmistesse klassidesse (arvulised näitajad on toodud treening-korpuse vigade põhjal):
1. Osalausepiiride valesti määramisest tingitud vead: 16. Näiteks:
(6) selle taga on saad aru selline lähenemine
Sõnavormilt *lähenemine* eemaldati aluse märgend, sest samas osalauses on ainsuse 2. pöördes verb.
2. Tundmatu süntaktiline funktsioon: 12.
Suulises kõnes on palju poolikuid, väljajättelisi või vigaseid lauseid, mistõttu ei ole võimalik kõigi sõnade süntaktilisi funktsioone inimesel määrata. Tundmatud funktsioonid märgendatakse @T-märgendiga. Samas ei ole võimalik koostada reegleid, et üks või teine sõnavorm on ilmselt määratlemata süntaktilise funktsiooniga. Analüsaator lisab sõnavormile morfoloogilise info ja konteksti põhjal kõik võimalikud

märgendid ning hakkab siis süntaktilisi kitsendusi rakendades neid ükshaaval eemaldama. Kui lause on poolik, siis pole piisavalt kontekstiinfot ning sõnale võib jääda mitu märgendit. Võib olla ka juhuseid, et reeglite põhjal üritatakse eemaldada kõiki märgendeid, kuid analüsaator ei luba kunagi eemaldada viimast. Nii võivad mittegrammatilises lauses olla sõnadel ka üsna juhuslikud süntaktilised funktsioonid. Näide 7 on pealeräkimise analüüsist. Nurksulgudes on käsitsi määratud märgendid, @-sümboliga aga tegelikult leitud. Sõnavorm *nad* jäi mitmeseks aluse ja sihitise vahel, käsitsi oli ta analüüsitud kui tundmatu süntaktilise funktsiooniga, sest kontekst, mille põhjal selgub tegelik funktsioon, paikneb järgmises kõnevoorus.

(7) A: et [J] @J nad [T] @SUBJ @OBJ
B: mhmh [B] @B
A: sobivad [FMV] @FMV kätte [ADVL] @ADVL

Näites 8 on tundmatu funktsiooniga sõnal juhuslik märgend

(8) Pöidla [T] @NN> ja [J] @J kõik [OBJ] @OBJ on [+FCV] @+FCV tehtud [-FMV] @+FMV

3. Omadussõna nimisõna rollis: 9.

Omadussõnad võivad elliptilistes lausetes esineda aluse või sihitise rollis (näide 9). Seda tüüpi vigu esineb ka kirjaliku keele analüüsil, kuid harvem.

(9) Ma [SUBJ] @SUBJ ostaks [+FMV] @+FMV selle [NN>] @OBJ maasikamaitselise [OBJ] @ADVL

4. Varasem vale analüüs: 5.

Mõni varem rakendatud reegel on analüüsinud mõnda sõnavormi valesti ning see vigane analüüs põhjustab ka naabersõnade vigast analüüsi.

5. Kordused: 3 (aluse kordamisel rikutakse unikaalsuse printsiipi).

(10) noh [B] @B se [SUBJ] @NN> see [SUBJ] @SUBJ on [FMV] @FMV tähtis [PRD] @PRD

6. Muu: 14 (võivad leiduda ka kirjaliku keele tekstis).

4.4. Mitmesused

Kui võrrelda kirjaliku keele ja suulise kõne allesjäänud mitmesuste klasse, siis ilmneb, et nende arvuline järjestus on erinev. Kirjaliku keele automaatselt analüüsitud tekstides oli domineeriv mitmesus määruse ja järeltäiendi vahel, kolm korda harvem esines sihitise ja eestäiendi, määruse ja eestäiendi ning aluse ja sihitise vahelist mitmesust. Suulise kõne tekstides põhjustavad arvukaimat mitmesust infiniitsed verbivormid, millele on jäänud nii öeldise kui ka määruse märgendid. Järgnevad määrus ja alus, määrus ja järeltäiend, alus ja sihitis, alus ja öeldistäide.

5. Mitteladususued

Suulise kõne süntaktilisel analüüsil tuleb hakkama saada mitmete suulisele kõnele omaste nähtustega, mida üldiselt nimetatakse mitteladususteks (ingl *disfluency*). Mitteladususteks peetakse näiteks kordusi, parandusi, poolikuks jäänud lausungeid. Eelpool nägime, et mitte kõik suulise kõne süntaktilisel analüüsil ilmnevad probleemid pole reeglitega lahendatavad. Üheks võimaluseks mitteladususte gа hakkama saada on eeltöötlemisetapis need märgendada, mille käigus ebagrammatilised lausungid tehakse süntaksianalüsaatori jaoks grammatilisteks, sageli nimetatakse seda protsessi ka normaliseerimiseks (ingl *normalization*). Mitteladususi on märgendatud näiteks suulise kõne korpuses Switchboard (Meteer jt 1995) ja ICE-GB (Meyer 2002: 96).

Switchboardi korpuses on mitteladususte märgendamiseks välja töötatud spetsiaalne märgendamisskeem, mis on võetud ka eesti suulise kõne märgendamise aluseks. Kasutusel olevad märgendid on ära toodud tabelis 2.

Märgend	*Seletus*	*Näide*
{D …}	diskursusemarker	**{D** nagu**}**; **{D** noh**}**
{F …}	täidetud paus	**{F** ee**}**; **{F** õ**}**
{B …}	hingamine	**{B** hh**}**
{A …}	raskesti analüüsitav	meil kül `präegu sin `kohapeal **{A** meil**}** `sellist vari`anti ei `paista. /
[RE … + …]	kordus	**[RE** nii + nii**]**
[RP … + …]	parandus	**[RP** kuidas + kas**]** **[RP** selli- + sellist**]**
/	lausung	H: tere, ma `sooviksin saada `infot õppe`laenu kohta. /
-/	lõpetamata lausung	V:siis saate sealt **-/** minu=arust ´dekanaat väljastab niisugused ´tõendid **/**
--	lausung jätkub	H: ma tahaksin tellida teie kataloogist seda ´juuksehooldusvahendit, -- V: [jaa] / H: -- [Blisaana.] /

Tabel 2: Mitteladususte analüüsil kasutatavad märgendid, nende seletus ja näited.

Vastavate märgenditega on analüüsitud 16 dialoogi (5631 sõna), kuid eelmärgendatud teksti pole veel jõutud automaatselt süntaktiliselt analüüsida, et hinnata selle skeemi rakendatavust tegelikkuses.

Märgendamisel tekkinud raskusi:

1. Lausungid ei lõpe alati intonatsiooni langusega, nt

H: ma:=ei=olnd `Tartus / mul jäi mine`mata=ja=nüüd ma [tahaks] uuesti `aega võtta/

2. Paranduse algus ja lõpp pole alati üheselt määratav. Sageli on raske vahet teha, kas eelnev üksus jäeti pooleli ja alustati uut või on tegu parandamisega, nt

H: **[RE** et + et**] -/ [RP** kas te k- + mille põhjal te**] {D** nagu**}** ´otsustate seda. /

Kuid on ka väga selgeid näiteid sellest, kus üks lausung jääb pooleli ja alustatakse uuega, nt

ma=i=´oska **{D** nagu**} {D** nimodi**}** kohe täpselt ´öelda et **{D** noh**}** selline: -/ **{B** .hh**}**
{Fm**}** ´tegemist on ühe üliõpilasorganisatsiooni ´aastapäevaga. /

3. Pealeräägitud kõne tulemuseks on palju poolelijäänud lausungeid, mis iseseisvana ei kanna endas just palju informatsiooni, seepärast on analüüsis lubatud minna üle kõnevooru piiri, et tekiks lauseline tervik, nt

V: `nii on väga `raske $ teile anda `infot, $ **{D** noh**}** =et see on [väga] --
H: **{D** [mhmh]**} /**
V: -- erineva `hinnaklassiga, erineva `pikkusega ja igal `maal on **{D** noh**}** omad ekskursi`oonid, oma `vaata[mis]väärsused. /

Vaatamata sellele, et ka käsitsi märgendamisel tekib küsitavusi, on selline märgendamine siiski suureks abiks. Kui vaadata eespool välja toodud automaatsel analüüsil tekkinud veatüüpe, siis mitteladususte märgendamine aitab osalt kindlasti lahendada lausungipiiride leidmise, poolelijäänud lausungite, paranduste ja kordustega seotud probleeme.

6. Järeldused ja tulevikuplaanid

Kirjaliku keele jaoks loodud süntaksianalüsaatori kohandamine suulisele kõnele osutus kergemaks ülesandeks kui algul loodetud. Kõige olulisem oli lahendada osalausepiiride määramise probleem, süntaktilised kitsendused vajasid vaid vähest muutmist. Üllatuslikult olid suulise keele süntaksianalüsaatori tulemused paremadki kui kirjaliku keele korral. Sellele on kaks põhjendust. 1) Suulises kõnes kasutatakse palju partikleid ja adverbe, mille süntaktiline analüüs on triviaalne. 2) Lausungid on lühemad.

Meie hinnangul on kitsenduste grammatika sobiv formalism suulise kõne süntaktiliseks analüüsiks: 1) analüüs on pindmine ning ei näita isegi põhja ja laiendi vahelist seost ilmutatult, see võimaldab ka valesti ühilduvad sõnavormid mõningatel juhtudel õigesti analüüsida; 2) sõnad, mida ei õnnestunud analüüsida, jäetakse mitmese analüüsiga, mis vähendab vigade hulka.

Väga raske on hinnata, kui lihtne oleks kohandada ESTKG morfoloogilist ühestajat suulisele kõnele. E. Bick (1998) jõudis oma katsetes järeldusele, et portugali keele kitsenduste grammatika ühestamise reeglite muutmine andis parema tulemuse kui süntaktiliste kitsenduste modifitseerimine, kuna morfoloogiline ühestamine vajab väiksemat konteksti. Sama ei pruugi kehtida eesti keele korral, keeled on selleks liiga erinevad. Loodetavasti edaspidised eksperimendid annavad sellele küsimusele peagi vastuse.

Töö käigus selgus, et kitsenduste grammatika ei sobi mõnede suulisele kõnele iseloomulike nähtuste märgendamiseks nagu kordused, valestardid ja eneseparandused. Seda sorti märgendus peaks olema tehtud juba enne süntaktilist analüüsi kasutades mõnda teist formalismi. Selle olemasolu hõlbustaks süntaksianalüsaatori tööd.

Me ei ole veel välja töötanud meetodit suulise kõne esitamiseks süntaksipuuna. Võib-olla oleks mõttekas kasutada sama lähenemist nagu poolautomaatsel kirjakeele puudepanga Arborest[2] loomisel (Bick jt. 2004).

Kirjandus

Bick, Eckhard 1998. Tagging Speech Data - Constraint Grammar Analysis of Spoken Portuguese, *Proceedings of the 17th Scandinavian Conference of Linguistics.* Odense.

Bick, Eckhard, Heli Uibo, Kaili Müürisep 2004. Arborest - a VISL-Style Treebank Derived from Estonian Constraint Grammar Corpus. *Proceedings of the Third Workshop on Treebanks and Linguistic Theories (TLT 2004).* Tübingen, Germany, Dec 10-11, 2004.

Erelt, Mati, Reet Kasik, Helle Metslang, Henno Rajandi, Kristiina Ross, Henn Saari, Kaja Tael, Silvi Vare 1993. *Eesti keele grammatika. II Süntaks.* Tallinn: Eesti TA Keele ja Kirjanduse Instituut.

Hennoste, Tiit 2002. Suulise kõne uurimine ja sõnaliigi probleemid. *Teoreetiline keeleteadus Eestis.* Tartu Ülikooli üldkeeleteaduse õppetooli toimetised 4. Tartu. 56-73.

Hennoste, Tiit, Mare Koit, Andriela Rääbis, Krista Strandson, Maret Valdisoo, Evelin Vutt 2003. Developing a Typology of Dialogue Acts: Tagging Estonian Dialogue Corpus. *DiaBruck 2003. Proceedings of the 7th Workshop on the Semantics and*

[2] http://corp.hum.sdu.dk/arborest.html

Pragmatics of Dialogue.(Eds) I. Kruijff-Korbayová, C. Kosny Saarland University, Saarbrücken , pp. 181-182

Hennoste, Tiit; Liina Lindström, Andriela Rääbis, Piret Toomet, Riina Vellerind 2000. Tartu University Corpus of Spoken Estonian. – T. Seilenthal, A. Nurk, T. Palo (eds.). *Congressus Nonus Internationalis Fenno-Ugristarum* 7.-13. 8. 2000. Pars IV. Dissertationes sectionum: Linguistica I, Tartu. 345-351.

Karlsson, Fred; Arto Anttila, Juha Heikkilä, Atro Voutilainen 1995. *Constraint Grammar: a Language-Independent System for Parsing Unrestricted Text.* Mouton de Gruyter.

Meteer, Marie; Ann Taylor, Robert MacIntyre, Rukmini Iyer 1995. *Dysfluency Annotation Stylebook for the Switchboard Corpus.* Linguistic Data Consortium. www.ldc.upenn.edu/Catalog/CatalogList/LDC99T42/DFLGUIDE.PS

Meyer, Charles F. 2002. *English Corpus Linguistics: An Introduction.* Cambridge University Press.

Roosmaa, Tiit, Mare Koit, Kadri Muischnek, Kaili Müürisep, Tiina Puolakainen, Heli Uibo 2003.. Eesti keele arvutigrammatika: mis on tehtud ja kuidas edasi? *Keel ja Kirjandus* nr 3, lk 192-209.